ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/220605213

Machine Ethics: Creating an Ethical Intelligent Agent.

Article *in* AI Magazine · December 2007

Source: DBLP

CITATIONS
399

READS
19,719

2 authors:

Michael Anderson
University of Hartford
69 PUBLICATIONS **2,394** CITATIONS
SEE PROFILE

Susan Leigh Anderson
University of Connecticut
64 PUBLICATIONS **2,331** CITATIONS
SEE PROFILE

All content following this page was uploaded by Michael Anderson on 02 June 2014.

The user has requested enhancement of the downloaded file.

AI Magazine Volume 28 Number 4 (2007) (© AAAI)

Machine Ethics: Creating an Ethical Intelligent Agent

Michael Anderson and Susan Leigh Anderson

■ The newly emerging field of machine ethics (Anderson and Anderson 2006) is concerned with adding an ethical dimension to machines. Unlike computer ethics—which has traditionally focused on ethical issues surrounding humans' use of machines—machine ethics is concerned with ensuring that the behavior of machines toward human users, and perhaps other machines as well, is ethically acceptable. In this article we discuss the importance of machine ethics, the need for machines that represent ethical principles explicitly, and the challenges facing those working on machine ethics. We also give an example of current research in the field that shows that it is possible, at least in a limited domain, for a machine to abstract an ethical principle from examples of correct ethical judgments and use that principle to guide its own behavior.

The ultimate goal of machine ethics, we believe, is to create a machine that *itself* follows an ideal ethical principle or set of principles; that is to say, it is guided by this principle or these principles in decisions it makes about possible courses of action it could take. We need to make a distinction between what James Moor has called an "*implicit ethical agent*" and an "*explicit ethical agent*" (Moor 2006). According to Moor, a machine that is an implicit ethical agent is one that has been programmed to behave ethically, or at least avoid unethical behavior, without an explicit representation of ethical principles. It is constrained in its behavior by its designer who is following ethical principles. A machine that is an explicit ethical agent, on the other hand, is able to calculate the best action in ethical dilemmas using ethical principles. It can "represent ethics explicitly and then operate effectively on the basis of this knowledge." Using Moor's terminology, most of those working on machine ethics would say that the ultimate goal is to create a machine that is an explicit ethical agent.

We are, here, primarily concerned with the ethical decision making itself, rather than how a machine would gather the information needed to make the decision and incorporate it into its general behavior. It is important to see this as a separate and considerable challenge. It is separate because having all the information and facility in the world won't, by itself, generate ethical behavior in a machine. One needs to turn to the branch of philosophy that is concerned with ethics for insight into what is considered to be ethically acceptable behavior. It is a considerable challenge because, even among experts, ethics has not been completely codified. It is a field that is still evolving. We shall argue that one of the advantages of working on machine ethics is that it might lead to breakthroughs in ethical theory, since machines are well-suited for testing the results of consistently following a particular ethical theory.

One other point should be made in introducing the subject of machine ethics. Ethics can be seen as both easy and hard. It appears easy because we all make ethical decisions on a daily basis. But that doesn't mean that we are all experts in ethics. It is a field that requires much study and experience. AI researchers must have respect for the expertise of ethicists just as ethicists must appreciate the expertise of AI researchers. Machine ethics is an inherently interdisciplinary field.

Copyright © 2007, American Association for Artificial Intelligence. All rights reserved. ISSN 0738-4602

The Importance of Machine Ethics

Why is the field of machine ethics important? There are at least three reasons that can be given. First, there are ethical ramifications to what machines currently do and are projected to do in the future. To neglect this aspect of machine behavior could have serious repercussions. South Korea has recently mustered more than 30 companies and 1000 scientists to the end of putting "a robot in every home by 2010" (Onishi 2006). DARPA's grand challenge to have a vehicle drive itself across 132 miles of desert terrain has been met, and a new grand challenge is in the works that will have vehicles maneuvering in an urban setting. The United States Army's Future Combat Systems program is developing armed robotic vehicles that will support ground troops with "direct-fire" and antitank weapons. From family cars that drive themselves and machines that discharge our daily chores with little or no assistance from us, to fully autonomous robotic entities that will begin to challenge our notions of the very nature of intelligence, it is clear that machines such as these will be capable of causing harm to human beings unless this is prevented by adding an ethical component to them.

Second, it could be argued that humans' fear of the possibility of autonomous intelligent machines stems from their concern about whether these machines will behave ethically, so the future of AI may be at stake. Whether society allows AI researchers to develop anything like autonomous intelligent machines may hinge on whether they are able to build in safeguards against unethical behavior. From the murderous robot uprising in the 1920 play *R.U.R.* (Capek 1921) and the deadly coup d'état perpetrated by the HAL 9000 computer in *2001: A Space Odyssey* (Clarke 1968), to *The Matrix* virtual reality simulation for the pacification and subjugation of human beings by machines, popular culture is rife with images of machines devoid of any ethical code mistreating their makers. In his widely circulated treatise, "Why the future doesn't need us," Bill Joy (2000) argues that the only antidote to such fates and worse is to "relinquish dangerous technologies." We believe that machine ethics research may offer a viable, more realistic solution.

Finally, we believe that it's possible that research in machine ethics will advance the study of ethical theory. Ethics, by its very nature, is the most practical branch of philosophy. It is concerned with how agents ought to behave when faced with ethical dilemmas. Despite the obvious applied nature of the field of ethics, too often work in ethical theory is done with little thought to actual application. When examples are discussed, they are typically artificial examples. Research in machine ethics has the potential to discover problems with current theories, perhaps even leading to the development of better theories, as AI researchers force scrutiny of the details involved in actually applying an ethical theory to particular cases. As Daniel Dennett (2006) recently stated, AI "makes philosophy honest." Ethics must be made computable in order to make it clear exactly how agents ought to behave in ethical dilemmas.

An exception to the general rule that ethicists don't spend enough time discussing actual cases occurs in the field of biomedical ethics, a field that has arisen out of a need to resolve pressing problems faced by health-care workers, insurers, hospital ethics boards, and biomedical researchers. As a result of there having been more discussion of actual cases in the field of biomedical ethics, a consensus is beginning to emerge as to how to evaluate ethical dilemmas in this domain, leading to the ethically correct action in many dilemmas. A reason there might be more of a consensus in this domain than in others is that in the area of biomedical ethics there is an ethically defensible goal (the best possible health of the patient), whereas in other areas (such as business and law) the goal may not be ethically defensible (make as much money as possible, serve the client's interest even if he or she is guilty of an offense or doesn't deserve a settlement) and ethics enters the picture as a limiting factor (the goal must be achieved within certain ethical boundaries).

AI researchers working with ethicists might find it helpful to begin with this domain, discovering a general approach to computing ethics that not only works in this domain, but could be applied to other domains as well.

Explicit Ethical Machines

It does seem clear, to those who have thought about the issue, that some sort of safeguard should be in place to prevent unethical machine behavior (and that work in this area may provide benefits for the study of ethical theory as well). This shows the need for creating at least *implicit* ethical machines; but why must we create *explicit* ethical machines, which would seem to be a much greater (perhaps even an impossible) challenge for AI researchers? Furthermore, many fear handing over the job of ethical over-

seer to machines themselves. How could we feel confident that a machine would make the right decision in situations that were not anticipated? Finally, what if the machine starts out behaving in an ethical fashion but then morphs into one that decides to behave unethically in order to secure advantages for itself?

On the need for *explicit,* rather than just *implicit,* ethical machines: What is critical in the "explicit ethical agent" versus "implicit ethical agent" distinction, in our view, lies not only in who is making the ethical judgments (the machine versus the human programmer), but also in the ability to *justify* ethical judgments that only an explicit representation of ethical principles allows. An explicit ethical agent is able to explain why a particular action is either right or wrong by appealing to an ethical principle. A machine that has learned, or been programmed, to make correct ethical judgments, but does not have principles to which it can appeal to justify or explain its judgments, is lacking something essential to being accepted as an ethical agent. Immanuel Kant (1785) made a similar point when he distinguished between an agent that acts from a sense of duty (consciously following an ethical principle), rather than merely in *accordance* with duty, having praise only for the former.

If we believe that machines could play a role in improving the lives of human beings—that this is a worthy goal of AI research—then, since it is likely that there will be ethical ramifications to their behavior, we must feel confident that these machines will act in a way that is ethically acceptable. It will be essential that they be able to justify their actions by appealing to acceptable ethical principles that they are following, in order to satisfy humans who will question their ability to act ethically. The ethical component of machines that affect humans' lives must be transparent, and principles that seem reasonable to human beings provide that transparency. Furthermore, the concern about how machines will behave in situations that were not anticipated also supports the need for explicit ethical machines. The virtue of having principles to follow, rather than being programmed in an ad hoc fashion to behave correctly in specific situations, is that it allows machines to have a way to determine the ethically correct action in new situations, even in new domains. Finally, Marcello Guarini (2006), who is working on a neural network model of machine ethics, where there is a predisposition to eliminate principles, argues that principles seem to play an important role in *revising* ethical beliefs, which is essential to ethical agency. He contends, for instance, that they are necessary to discern morally relevant differences in similar cases.

The concern that machines that start out behaving ethically will end up behaving unethically, perhaps favoring their own interests, may stem from fears derived from legitimate concerns about *human* behavior. Most human beings are far from ideal models of ethical agents, despite having been taught ethical principles; and humans do, in particular, tend to favor themselves. Machines, though, might have an advantage over human beings in terms of behaving ethically. As Eric Dietrich (2006) has recently argued, human beings, as biological entities in competition with others, may have evolved into beings with a genetic predisposition toward unethical behavior as a survival mechanism. Now, though, we have the chance to create entities that lack this predisposition, entities that might even inspire us to behave more ethically. Consider, for example, Andrew, the robot hero of Isaac Asimov's story "The Bicentennial Man" (1976), who was far more ethical than the humans with whom he came in contact. Dietrich maintained that the machines we fashion to have the good qualities of human beings, and that also follow principles derived from ethicists who are the exception to the general rule of unethical human beings, could be viewed as "humans 2.0"—a better version of human beings.

This may not completely satisfy those who are concerned about a future in which human beings share an existence with intelligent, autonomous machines. We face a choice, then, between allowing AI researchers to continue in their quest to develop intelligent, autonomous machines—which will have to involve adding an ethical component to them—or stifling this research. The likely benefits and possible harms of each option will have to be weighed. In any case, there are certain benefits to continuing to work on machine ethics. It is important to find a clear, objective basis for ethics—making ethics in principle computable—if only to rein in unethical *human* behavior; and AI researchers, working with ethicists, have a better chance of achieving breakthroughs in ethical theory than theoretical ethicists working alone. There is also the possibility that society would not be able to prevent some researchers from continuing to develop intelligent, autonomous machines, even if society decides that it is too dangerous to support such work. If this research should be successful, it will be important that we have ethical principles that we insist should be incorporated into such machines. The one thing that society should fear more than sharing an existence with intel-

ligent, autonomous machines is sharing an existence with machines like these without an ethical component.

Challenges Facing Those Working on Machine Ethics

The challenges facing those working on machine ethics can be divided into two main categories: philosophical concerns about the feasibility of computing ethics and challenges from the AI perspective. In the first category, we need to ask whether ethics is the sort of thing that can be computed. One well-known ethical theory that supports an affirmative answer to this question is "act utilitarianism." According to this teleological theory (a theory that maintains that the rightness and wrongness of actions is determined entirely by the consequences of the actions) that act is right which, of all the actions open to the agent, is likely to result in the greatest net good consequences, taking all those affected by the action equally into account. Essentially, as Jeremy Bentham (1781) long ago pointed out, the theory involves performing "moral arithmetic."

Of course, before doing the arithmetic, one needs to know what counts as a "good" and "bad" consequence. The most popular version of act utilitarianism—hedonistic act utilitarianism—would have us consider the pleasure and displeasure that those affected by each possible action are likely to receive. And, as Bentham pointed out, we would probably need some sort of scale to account for such things as the intensity and duration of the pleasure or displeasure that each individual affected is likely to receive. This is information that a human being would need to have as well to follow the theory. Getting this information has been and will continue to be a challenge for artificial intelligence research in general, but it can be separated from the challenge of computing the ethically correct action, given this information. With the requisite information, a machine could be developed that is just as able to follow the theory as a human being.

Hedonistic act utilitarianism can be implemented in a straightforward manner. The algorithm is to compute the best action, that which derives the greatest net pleasure, from all alternative actions. It requires as input the number of people affected and, for each person, the intensity of the pleasure/displeasure (for example, on a scale of 2 to –2), the duration of the pleasure/displeasure (for example, in days), and the probability that this pleasure or displeasure will occur, for each possible action. For each person, the algorithm computes the product of the intensity, the duration, and the probability, to obtain the net pleasure for that person. It then adds the individual net pleasures to obtain the total net pleasure:

Total net pleasure = $\sum$ (intensity × duration × probability) for each affected individual

This computation would be performed for each alternative action. The action with the highest total net pleasure is the right action (Anderson, Anderson, and Armen 2005b).

A machine might very well have an advantage over a human being in following the theory of act utilitarianism for several reasons: First, human beings tend not to do the arithmetic strictly, but just estimate that a certain action is likely to result in the greatest net good consequences, and so a human being might make a mistake, whereas such error by a machine would be less likely. Second, as has already been noted, human beings tend toward partiality (favoring themselves, or those near and dear to them, over others who might be affected by their actions or inactions), whereas an impartial machine could be devised. Since the theory of act utilitarianism was developed to introduce objectivity into ethical decision making, this is important. Third, humans tend not to consider all of the possible actions that they could perform in a particular situation, whereas a more thorough machine could be developed. Imagine a machine that acts as an advisor to human beings and "thinks" like an act utilitarian. It will prompt the human user to consider alternative actions that might result in greater net good consequences than the action the human being is considering doing, and it will prompt the human to consider the effects of each of those actions on all those affected. Finally, for some individuals' actions—actions of the president of the United States or the CEO of a large international corporation—their impact can be so great that the calculation of the greatest net pleasure may be very time consuming, and the speed of today's machines gives them an advantage.

We conclude, then, that machines can follow the theory of act utilitarianism at least as well as human beings and, perhaps, even better, *given the data that human beings would need,* as well, to follow the theory. The theory of act utilitarianism has, however, been questioned as not entirely agreeing with intuition. It is certainly a good starting point in programming a machine to be ethically sensitive—it would probably be more ethically sensitive than many human beings—but, perhaps, a better ethical theory can be used.

Critics of act utilitarianism have pointed out that it can violate human beings' rights, sacrificing one person for the greater net good. It can also conflict with our notion of justice—what people deserve—because the rightness and wrongness of actions is determined entirely by the future consequences of actions, whereas what people deserve is a result of past behavior. A deontological approach to ethics (where the rightness and wrongness of actions depends on something other than the consequences), such as Kant's categorical imperative, can emphasize the importance of rights and justice, but this approach can be accused of ignoring consequences. We believe, along with W. D. Ross (1930), that the best approach to ethical theory is one that combines elements of both teleological and deontological theories. A theory with several prima facie duties (obligations that we should try to satisfy, but which can be overridden on occasion by stronger obligations)—some concerned with the consequences of actions and others concerned with justice and rights—better acknowledges the complexities of ethical decision making than a single absolute duty theory. This approach

has one major drawback, however. It needs to be supplemented with a decision procedure for cases where the prima facie duties give conflicting advice. This is a problem that we have worked on and will be discussed later on.

Among those who maintain that ethics cannot be computed, there are those who question the action-based approach to ethics that is assumed by defenders of act utilitarianism, Kant's categorical imperative, and other well-known ethical theories. According to the "virtue" approach to ethics, we should not be asking what we ought to do in ethical dilemmas, but rather what sort of persons we should be. We should be talking about the sort of qualities—virtues—that a person should possess; actions should be viewed as secondary. Given that we are concerned only with the actions of machines, it is appropriate, however, that we adopt the action-based approach to ethical theory and focus on the sort of principles that machines should follow in order to behave ethically.

Another philosophical concern with the machine ethics project is whether machines are the type of entities that can behave ethically. It is commonly thought that an entity must be capable of acting intentionally, which requires that it be conscious, and that it have free will, in order to be a moral agent. Many would, also, add that sentience or emotionality is important, since only a being that has feelings would be capable of appreciating the feelings of others, a critical factor in the moral assessment of possible actions that could be performed in a given situation. Since many doubt that machines will ever be conscious, have free will, or emotions, this would seem to rule them out as being moral agents.

This type of objection, however, shows that the critic has not recognized an important distinction between performing the morally correct action in a given situation, including being able to justify it by appealing to an acceptable ethical principle, and being held morally responsible for the action. Yes, intentionality and free will in some sense are necessary to hold a being morally responsible for its actions, and it would be difficult to establish that a machine possesses these qualities; but neither attribute is necessary to do the morally correct action in an ethical dilemma and justify it. All that is required is that the machine act in a way that conforms with what would be considered to be the morally correct action in that situation and be able to justify its action by citing an acceptable ethical principle that it is following (S. L. Anderson 1995).

The connection between emotionality and being able to perform the morally correct action in an ethical dilemma is more complicated. Certainly one has to be sensitive to the suffering of others to act morally. This, for human beings, means that one must have empathy, which, in turn, requires that one have experienced similar emotions oneself. It is not clear, however, that a machine could not be trained to take into account the suffering of others in calculating how it should behave in an ethical dilemma, without having emotions itself. It is important to recognize, furthermore, that having emotions can actually interfere with a being's ability to determine, and perform, the right action in an ethical dilemma. Humans are prone to getting "carried away" by their emotions to the point where they are incapable of following moral principles. So emotionality can even be viewed as a weakness of human beings that often prevents them from doing the "right thing."

The necessity of emotions in rational decision making in computers has been championed by Rosalind Picard (1997), citing the work of Damasio (1994), which concludes that human beings lacking emotion repeatedly make the same bad decisions or are unable to make decisions in due time. We believe that, although evolution may have taken this circuitous path to decision making in human beings, irrational control of rational processes is not a necessary condition for all rational systems—in particular, those specifically designed to learn from errors, heuristically prune search spaces, and make decisions in the face of bounded time and knowledge.

A final philosophical concern with the feasibility of computing ethics has to do with whether there is a single correct action in ethical dilemmas. Many believe that ethics is relative either to the society in which one lives—"when in Rome, one should do what Romans do"—or, a more extreme version of relativism, to individuals—whatever you think is right is right for you. Most ethicists reject ethical relativism (for example, see Mappes and DeGrazia [2001, p. 38] and Gazzaniga [2006, p. 178]), in both forms, primarily because this view entails that one cannot criticize the actions of societies, as long as they are approved by the majority in those societies, or individuals who act according to their beliefs, no matter how heinous they are. There certainly do seem to be actions that experts in ethics, and most of us, believe are absolutely wrong (for example, torturing a baby and slavery), even if there are societies, or individuals, who approve of the actions. Against those who say that ethical relativism is a more tolerant view than ethical absolutism, it has been pointed out that ethical relativists cannot say that anything is absolutely good—even tolerance (Pojman [1996, p. 13]).

What defenders of ethical relativism may be recognizing—that causes them to support this view—are two truths, neither of which entails the acceptance of ethical relativism: (1) Different societies have their own customs that we must acknowledge, and (2) there are difficult ethical issues about which even experts in ethics cannot agree, at the present time, on the ethically correct action. Concerning the first truth, we must distinguish between an ethical issue and customs or practices that fall outside the area of ethical concern. Customs or practices that are not a matter of ethical concern can be respected, but in areas of ethical concern we should not be tolerant of unethical practices.

Concerning the second truth, that *some* ethical issues are difficult to resolve (for example, abortion)—and so, at this time, there may not be agreement by ethicists as to the correct action—it does not follow that all views on these issues are equally correct. It will take more time to resolve

these issues, but most ethicists believe that we should strive for a single correct position on these issues. What needs to happen is to see that a certain position follows from basic principles that all ethicists accept, or that a certain position is more consistent with other beliefs that they all accept.

From this last point, we should see that we may not be able to give machines principles that resolve *all* ethical disputes at this time. (Hopefully, the machine behavior that we are concerned about won't fall in too many of the disputed areas.) The implementation of ethics can't be more complete than is accepted ethical theory. Completeness is an ideal for which to strive but may not be possible at this time. The ethical theory, or framework for resolving ethical disputes, should allow for updates, as issues that once were considered contentious are resolved. What is more important than having a complete ethical theory to implement is to have one that is consistent. This is where machines may actually help to advance the study of ethical theory, by pointing out inconsistencies in the theory that one attempts to implement, forcing ethical theoreticians to resolve those inconsistencies.

Considering challenges from an AI perspective, foremost for the nascent field of machine ethics may be convincing the AI community of the necessity and advisability of incorporating ethical principles into machines. Some critics maintain that machine ethics is the stuff of science fiction—machines are not yet (and may never be) sophisticated enough to require ethical restraint. Others wonder who would deploy such systems given the possible liability involved. We contend that machines with a level of autonomy requiring ethical deliberation are here and both their number and level of autonomy are likely to increase. The liability already exists; machine ethics is necessary as a means to mitigate it. In the following section, we will detail a system that helps establish this claim.

Another challenge facing those concerned with machine ethics is how to proceed in such an inherently interdisciplinary endeavor. Artificial Intelligence researchers and philosophers, although generally on speaking terms, do not always hear what the other is saying. It is clear that, for substantive advancement of the field of machine ethics, both are going to have to listen to each other intently. AI researchers will need to admit their naiveté in the field of ethics and convince philosophers that there is a pressing need for their services; philosophers will need to be a bit more pragmatic than many are wont to be and make an effort to sharpen ethical theory in domains where machines will be active. Both will have to come to terms with this newly spawned relationship and, together, forge a common language and research methodology.

The machine ethics research agenda will involve testing the feasibility of a variety of approaches to capturing ethical reasoning, with differing ethical bases and implementation formalisms, and applying this reasoning in systems engaged in ethically sensitive activities. This research will investigate how to determine and represent ethical principles, incorporate ethical principles into a system's decision procedure, make ethical decisions with incomplete and uncertain knowledge, provide explanations for decisions made using ethical principles, and evaluate systems that act based upon ethical principles.

System implementation work is already underway. A range of machine-learning techniques are being employed in an attempt to codify ethical reasoning from examples of particular ethical dilemmas. As such, this work is based, to a greater or lesser degree, upon *casuistry*—the branch of applied ethics that, eschewing principle-based approaches to ethics, attempts to determine correct responses to new ethical dilemmas by drawing conclusions based on parallels with previous cases in which there is agreement concerning the correct response.

Rafal Rzepka and Kenji Araki (2005), at what might be considered the most extreme degree of casuistry, explore how statistics learned from examples of ethical intuition drawn from the full spectrum of the world wide web might be useful in furthering machine ethics. Working in the domain of safety assurance for household robots, they question whether machines should be obeying some set of rules decided by ethicists, concerned that these rules may not in fact be truly universal. They suggest that it might be safer to have machines "imitating millions, not a few," believing in such "democracy-dependent algorithms" because, they contend, "most people behave ethically without learning ethics." They propose an extension to their web-based knowledge discovery system GENTA (General Belief Retrieving Agent) that would search the web for opinions, usual behaviors, common consequences, and exceptions, by counting ethically relevant neighboring words and phrases, aligning these along a continuum from positive to negative behaviors, and subjecting this information to statistical analysis. They suggest that this analysis, in turn, would be helpful in the development of a sort of majority-rule ethics useful in guiding the behavior of autonomous systems. An important open question is whether users will be comfortable with such behavior or will, as might be expected, demand better than average ethical conduct from autonomous systems.

A neural network approach is offered by Marcello Guarini (2006). At what might be considered a less extreme degree of casuistry, particular actions concerning killing and allowing to die are classified as acceptable or unacceptable depending upon different motives and consequences. After training a simple recurrent network on a number of such cases, it is capable of providing plausible responses to a variety of previously unseen cases. This work attempts to shed light on the philosophical debate concerning *generalism* (principle-based approaches to moral reasoning) versus *particularism* (case-based approaches to moral reasoning). Guarini finds that, although some of the concerns pertaining to learning and generalizing from ethical dilemmas without resorting to principles can be mitigated with a neural network model of cognition, "important considerations suggest that it cannot be the whole story about moral reasoning—principles are needed." He argues that "to build an

artificially intelligent agent without the ability to question and revise its own initial instruction on cases is to assume a kind of moral and engineering perfection on the part of the designer." He argues, further, that such perfection is unlikely and principles seem to play an important role in the required subsequent revision—"at least some reflection in humans does appear to require the explicit representation or consultation of…rules," for instance, in discerning morally relevant differences in similar cases. Concerns about this approach are those attributable to neural networks in general, including oversensitivity to training cases and the inability to generate reasoned arguments for system responses.

Bruce McLaren (2003), in the spirit of a more pure form of casuistry, promotes a case-based reasoning approach (in the artificial intelligence sense) for developing systems that provide guidance in ethical dilemmas. His first such system, Truth-Teller, compares pairs of cases presenting ethical dilemmas about whether or not to tell the truth.

The Truth-Teller program marshals ethically relevant similarities and differences between two given cases from the perspective of the "truth teller" (that is, the person faced with the dilemma) and reports them to the user. In particular, it points out reasons for telling the truth (or not) that (1) apply to both cases, (2) apply more strongly in one case than another, or (3) apply to only one case.

The System for Intelligent Retrieval of Operationalized Cases and Codes (SIROCCO), McLaren's second program, leverages information concerning a new ethical dilemma to predict which previously stored principles and cases are relevant to it in the domain of professional engineering ethics. Cases are exhaustively formalized and this formalism is used to index similar cases in a database of formalized, previously solved cases that include principles used in their solution. SIROCCO's goal, given a new case to analyze, is "to provide the basic information with which a human reasoner … could answer an ethical question and then build an argument or rationale for that conclusion." SIROCCO is successful at retrieving relevant cases but performed beneath the level of an ethical review board presented with the same task. Deductive techniques, as well as any attempt at decision making, are eschewed by McLaren due to "the ill-defined nature of problem solving in ethics." Critics might contend that this "ill-defined nature" may not make problem solving in ethics completely indefinable, and attempts at just such a definition may be possible in constrained domains. Further, it might be argued that decisions offered by a system that are consistent with decisions made in previous cases have merit and will be useful to those seeking ethical advice.

We (Anderson, Anderson, and Armen 2006a) have developed a decision procedure for an ethical theory in a constrained domain that has multiple prima facie duties, using inductive logic programming (ILP) (Lavrec and Dzeroski 1997) to learn the relationships between these duties. In agreement with Marcello Guarini and Baruch Brody (1988) that casuistry alone is not sufficient, we begin with prima facie duties that often give conflicting advice in ethical dilemmas and then abstract a decision principle, when conflicts do arise, from cases of ethical dilemmas where ethicists are in agreement as to the correct action. We have adopted a multiple prima facie duty approach to ethical decision making because we believe it is more likely to capture the complexities of ethical decision making than a single, absolute duty ethical theory. In an attempt to develop a decision procedure for determining the ethically correct action when the duties give conflicting advice, we use ILP to abstract information leading to a general decision principle from ethical experts' intuitions about particular ethical dilemmas. A common criticism is whether the relatively straightforward representation scheme used to represent ethical dilemmas will be sufficient to represent a wider variety of cases in different domains.

Deontic logic's formalization of the notions of obligation, permission, and related concepts[1] make it a prime candidate as a language for the expression of machine ethics principles. Selmer Bringsjord, Konstantine Arkoudas, and Paul Bello (2006) show how formal logics of action, obligation, and permissibility might be used to incorporate a given set of ethical principles into the decision procedure of an autonomous system. They contend that such logics would allow for proofs establishing that (1) robots only take permissible actions, and (2) all actions that are obligatory for robots are actually performed by them, subject to ties and conflicts among available actions. They further argue that, while some may object to the wisdom of logic-based AI in general, they believe that in this case a logic-based approach is promising because one of the central issues in machine ethics is trust and "mechanized formal proofs are perhaps the single most effective tool at our disposal for establishing trust." Making no commitment as to the ethical content, their objective is to arrive at a methodology that maximizes the probability that an artificial intelligent agent behaves in a certifiably ethical fashion, subject to proof explainable in ordinary English. They propose a general methodology for implementing deontic logics in their logical framework, Athena, and illustrate the feasibility of this approach by encoding a natural deduction system for a deontic logic for reasoning about what agents ought to do. Concerns remain regarding the practical relevance of the formal logics they are investigating and efficiency issues in their implementation.

The work of Bringjord, Arkoudas, and Bello is based on research that investigates, from perspectives other than artificial intelligence, how deontic logic's concern with *what ought to be the case* might be extended to represent and reason about *what agents ought to do*. It has been argued that the implied assumption that the latter will simply follow from investigation of the former is not the case. In this context, John Horty (2001) proposes an extension of deontic logic, incorporating a formal theory of agency that describes what agents ought to do under various conditions over extended periods of time. In particular, he adapts preference ordering from deci-

sion theory to "both define optimal actions that an agent should perform and the propositions whose truth the agent should guarantee." This framework permits the uniform formalization of a variety of issues of ethical theory and, hence, facilitates the discussion of these issues.

Tom Powers (2006) assesses the feasibility of using deontic and default logics to implement Kant's categorical imperative:

> Act only according to that maxim whereby you can at the same time will that it should become a universal law… If contradiction and contrast arise, the action is rejected; if harmony and concord arise, it is accepted. From this comes the ability to take moral positions as a heuristic means. For we are social beings by nature, and what we do not accept in others, we cannot sincerely accept in ourselves.

Powers suggests that a machine might itself construct a theory of ethics by applying a universalization step to individual maxims, mapping them into the deontic categories of forbidden, permissible, or obligatory actions. Further, for consistency, these universalized maxims need to be tested for contradictions with an already established base of principles, and these contradictions resolved. Powers suggests, further, that such a system will require support from a theory of commonsense reasoning in which postulates must "survive the occasional defeat," thus producing a nonmonotonic theory whose implementation will require some form of default reasoning. It has been noted (Ganascia 2007) that answer set programming (ASP) (Baral 2003) may serve as an efficient formalism for modeling such ethical reasoning. An open question is what reason, other than temporal priority, can be given for keeping the whole set of prior maxims and disallowing a new contradictory one. Powers offers that "if we are to construe Kant's test as a way to build a set of maxims, we must establish rules of priority for accepting each additional maxim." The question remains as to what will constitute this moral epistemic commitment.

Creating a Machine That Is an Explicit Ethical Agent

To demonstrate the possibility of creating a machine that is an explicit ethical agent, we have attempted in our research to complete the following six steps:

Step One

We have adopted the prima facie duty approach to ethical theory, which, as we have argued, better reveals the complexity of ethical decision making than single, absolute duty theories. It incorporates the good aspects of the teleological and deontological approaches to ethics, while allowing for needed exceptions to adopting one or the other approach exclusively. It also has the advantage of being better able to adapt to the specific concerns of ethical dilemmas in different domains. There may be slightly different sets of prima facie duties for biomedical ethics, legal ethics, business ethics, and journalistic ethics, for example.

There are two well-known prima facie duty theories: Ross's theory, dealing with general ethical dilemmas, that has seven duties; and Beauchamp and Childress's four principles of biomedical ethics (1979) (three of which are derived from Ross's theory) that are intended to cover ethical dilemmas specific to the field of biomedicine. Because there is more agreement between ethicists working on biomedical ethics than in other areas, and because there are fewer duties, we decided to begin to develop our prima facie duty approach to computing ethics using Beauchamp and Childress's principles of biomedical ethics.

Beauchamp and Childress's principles of biomedical ethics include the principle of respect for autonomy that states that the health-care professional should not interfere with the effective exercise of patient autonomy. For a decision by a patient concerning his or her care to be considered fully autonomous, it must be intentional, based on sufficient understanding of his or her medical situation and the likely consequences of forgoing treatment, sufficiently free of external constraints (for example, pressure by others or external circumstances, such as a lack of funds) and sufficiently free of internal constraints (for example, pain or discomfort, the effects of medication, irrational fears, or values that are likely to change over time). The principle of nonmaleficence requires that the health-care professional not harm the patient, while the principle of beneficence states that the health-care professional should promote patient welfare. Finally, the principle of justice states that health-care services and burdens should be distributed in a just fashion.

Step Two

The domain we selected was medical ethics, consistent with our choice of prima facie duties, and, in particular, a representative type of ethical dilemma that involves three of the four principles of biomedical ethics: respect for autonomy, nonmaleficence, and beneficence. The type of dilemma is one that health-care workers often face: A health-care worker has recommended a particular treatment for her competent adult patient, and the patient has rejected that treatment option. Should the health-care worker try again to change the patient's mind or accept the patient's decision as final? The dilemma arises because, on the one hand, the health-care professional shouldn't challenge the patient's autonomy unnecessarily; on the other hand, the health-care worker may have concerns about why the patient is refusing the treatment.

In this type of dilemma, the options for the health-care worker are just two, either to accept the patient's decision or not, by trying again to change the patient's mind. For this proof of concept test of attempting to make a prima facie duty ethical theory computable, we have a single type of dilemma that encompasses a finite number of specific cases, just three duties, and only two possible actions in each case. We have abstracted, from a discussion of similar types of cases given by Buchanan and Brock (1989), the correct answers to the specific cases of the type of dilemma we consider. We have made the assumption that there is a consensus among bioethicists that these are the correct answers.

Step Three

The major philosophical problem with the prima facie duty approach to ethical decision making is the lack of a decision procedure when the duties give conflicting advice. What is needed, in our view, are ethical principles that balance the level of satisfaction or violation of these duties and an algorithm that takes case profiles and outputs that action that is consistent with these principles. A profile of an ethical dilemma consists of an ordered set of numbers for each of the possible actions that could be performed, where the numbers reflect whether particular duties are satisfied or violated and, if so, to what degree. John Rawls's "reflective equilibrium" (1951) approach to creating and refining ethical principles has inspired our solution to the problem of a lack of a decision procedure. We abstract a principle from the profiles of specific cases of ethical dilemmas where experts in ethics have clear intuitions about the correct action and then test the principle on other cases, refining the principle as needed.

The selection of the range of possible satisfaction or violation levels of a particular duty should, ideally, depend upon how many gradations are needed to distinguish between cases that are ethically distinguishable. Further, it is possible that new duties may need to be added in order to make distinctions between ethically distinguishable cases that would otherwise have the same profiles. There is a clear advantage to our approach to ethical decision making in that it can accommodate changes to the range of intensities of the satisfaction or violation of duties, as well as adding duties as needed.

Step Four

Implementing the algorithm for the theory required formulation of a principle to determine the correct action when the duties give conflicting advice. We developed a system (Anderson, Anderson, and Armen 2006a) that uses machine-learning techniques to abstract relationships between the prima facie duties from particular ethical dilemmas where there is an agreed-upon correct action. Our chosen type of dilemma, detailed previously, has only 18 possible cases (given a range of +2 to –2 for the level of satisfaction or violation of the duties) where, given the two possible actions, the first action supersedes the second (that is, was ethically preferable). Four of these cases were provided to the system as examples of when the target predicate (supersedes) is true. Four examples of when the target predicate is false were provided by simply reversing the order of the actions. The system discovered a principle that provides the correct answer for the remaining 14 positive cases, as verified by the consensus of ethicists.

ILP was used as the method of learning in this system. ILP is concerned with inductively learning relations represented as first-order Horn clauses (that is, universally quantified conjunctions of positive literals L_i implying a positive literal H: $H \leftarrow (L_1 \wedge \cdots \wedge L_n)$). ILP is used to learn the relation supersedes ($A1$, $A2$), which states that action $A1$ is preferred over action $A2$ in an ethical dilemma involving these choices. Actions are represented as ordered sets of integer values in the range of +2 to –2 where each value denotes the satisfaction (positive values) or violation (negative values) of each duty involved in that action. Clauses in the supersedes predicate are represented as disjunctions of lower bounds for differentials of these values.

ILP was chosen to learn this relation for a number of reasons. The potentially nonclassical relationships that might exist between prima facie duties are more likely to be expressible in the rich representation language provided by ILP than in less expressive representations. Further, the consistency of a hypothesis regarding the relationships between prima facie duties can be automatically confirmed across all cases when represented as Horn clauses. Finally, commonsense background knowledge regarding the supersedes relationship is more readily expressed and consulted in ILP's declarative representation language.

The object of training is to learn a new hypothesis that is, in relation to all input cases, complete and consistent. Defining a *positive* example as a case in which the first action supersedes the second action and a *negative* example as one in which this is not the case, a complete hypothesis is one that covers all positive cases, and a consistent hypothesis covers no negative cases. Negative training examples are generated from positive training examples by inverting the order of these actions, causing the first action to be the incorrect choice. The system starts with the most general hypothesis stating that all actions supersede each other and, thus, covers all positive and negative cases. The system is then provided with positive cases (and their negatives) and modifies its hypothesis, by adding or refining clauses, such that it covers given positive cases and does not cover given negative cases.

The decision principle that the system discovered can be stated as follows: A health-care worker should challenge a patient's decision if it isn't fully autonomous and there's either any violation of nonmaleficence or a severe violation of beneficence. Although, clearly, this rule is implicit in the judgments of the consensus of ethicists, to our knowledge this principle has never before been stated explicitly. Ethical theory has not yet advanced to the point where principles like this one—that correctly balance potentially conflicting duties with differing levels of satisfaction or violation—have been formulated. It is a significant result that machine-learning techniques can discover a principle such as this and help advance the field of ethics. We offer it as evidence that making the ethics more precise will permit machine-learning techniques to discover philosophically novel and interesting principles in ethics because the learning system is general enough that it can be used to learn relationships between any set of prima facie duties where there is a consensus among ethicists as to the correct answer in particular cases.

Once the principle was discovered, the needed decision procedure could be fashioned. Given a profile representing the satisfaction/violation levels of the duties involved in each possible action, values of corresponding

duties are subtracted (those of the second action from those of the first). The principle is then consulted to see if the resulting differentials satisfy any of its clauses. If so, the first action of the profile is deemed ethically preferable to the second.

Step Five

We have explored two prototype applications of the discovered principle governing Beauchamp and Childress's principles of biomedical ethics. In both prototypes, we created a program where a machine could use the principle to determine the correct answer in ethical dilemmas. The first, MedEthEx (Anderson, Anderson, and Armen 2006b), is a medical ethical advisor system; the second, EthEl, is a system in the domain of elder care that determines when a patient should be reminded to take medication and when a refusal to do so is serious enough to contact an overseer. EthEl is more autonomous than MedEthEx in that, whereas MedEthEx gives the ethically correct answer (that is, that which is consistent with its training) to a human user who will act on it or not, EthEl herself acts on what she determines to be the ethically correct action.

MedEthEx is an expert system that uses the discovered principle and decision procedure to give advice to a user faced with a case of the dilemma type previously described. In order to permit use by someone unfamiliar with the representation details required by the decision procedure, a user interface was developed that (1) asks ethically relevant questions of the user regarding the particular case at hand, (2) transforms the answers to these questions into the appropriate profiles, (3) sends these profiles to the decision procedure, (4) presents the answer provided by the decision procedure, and (5) provides a justification for this answer.[2]

The principle discovered can be used by other systems, as well, to provide ethical guidance for their actions. Our current research uses the principle to elicit ethically sensitive behavior from an elder-care system, EthEl, faced with a different but analogous ethical dilemma. EthEl must remind the patient to take his or her medication and decide when to accept a patient's refusal to take a medication that might prevent harm or provide benefit to the patient and when to notify an overseer. This dilemma is analogous to the original dilemma in that the same duties are involved (nonmaleficence, beneficence, and respect for autonomy) and "notifying the overseer" in the new dilemma corresponds to "trying again" in the original.

Machines are currently in use that face this dilemma.[3] The state of the art in these reminder systems entails providing "context-awareness" (that is, a characterization of the current situation of a person) to make reminders more efficient and natural. Unfortunately, this awareness does not include consideration of ethical duties that such a system should adhere to when interacting with its patient. In an ethically sensitive elder-care system, both the timing of reminders and responses to a patient's disregard of them should be tied to the duties involved. The system should challenge patient autonomy only when necessary, as well as minimize harm and loss of benefit to the patient. The principle discovered from the original dilemma can be used to achieve these goals by directing the system to remind the patient only at ethically justifiable times and notifying the overseer only when the harm or loss of benefit reaches a critical level.

In the implementation, EthEl receives input from an overseer (most likely a doctor), including: the prescribed time to take a medication, the maximum amount of harm that could occur if this medication is not taken (for example, none, some, or considerable), the number of hours it would take for this maximum harm to occur, the maximum amount of expected good to be derived from taking this medication, and the number of hours it would take for this benefit to be lost. The system then determines from this input the change in duty satisfaction and violation levels over time, a function of the maximum amount of harm or good and the number of hours for this effect to take place. This value is used to increment duty satisfaction and violation levels for the *remind* action and, when a patient disregards a reminder, the *notify* action. It is used to decrement *don't remind* and *don't notify* actions as well. A reminder is issued when, according to the principle, the duty satisfaction or violation levels have reached the point where reminding is ethically preferable to not reminding. Similarly, the overseer is notified when a patient has disregarded reminders to take medication and the duty satisfaction or violation levels have reached the point where notifying the overseer is ethically preferable to not notifying the overseer.

EthEl uses an ethical principle discovered by a machine to determine reminders and notifications in a way that is proportional to the amount of maximum harm to be avoided or good to be achieved by taking a particular medication, while not unnecessarily challenging a patient's autonomy. EthEl minimally satisfies the requirements of an explicit ethical agent (in a constrained domain), according to Jim Moor's definition of the term: A machine that is able to calculate the best action in ethical dilemmas using an ethical principle, as opposed to having been programmed to behave ethically, where the programmer is following an ethical principle.

Step Six

As a possible means of assessing the morality of a system's behavior, Colin Allen, G. Varner, and J. Zinser (2000) describe a variant of the test Alan Turing (1950) suggested as a means to determine the intelligence of a machine that bypassed disagreements about the definition of intelligence. Their proposed "comparative moral Turing test" (cMTT) bypasses disagreement concerning definitions of ethical behavior as well as the requirement that a machine have the ability to articulate its decisions: an evaluator assesses the comparative morality of pairs of descriptions of morally significant behavior where one describes the actions of a human being in an ethical dilemma and the other the actions of a machine faced with the same dilemma. If the machine is not identified as the less moral member of the pair significantly more often than the human, then it has passed the test.

They point out, though, that human behavior is typically far from being morally ideal and a machine that passed the cMTT might still fall far below the high ethical standards to which we would probably desire a machine to be held. This legitimate concern suggests to us that, instead of comparing the machine's behavior in a particular dilemma against typical human behavior, the comparison ought to be made with behavior recommended by a trained ethicist faced with the same dilemma. We also believe that the principles used to justify the decisions that are reached by both the machine and ethicist should be made transparent and compared.

We plan to devise and carry out a moral Turing test of this type in future work, but we have had some assessment of the work that we have done to date. The decision principle that was discovered in MedEthEx, and used by EthEl, is supported by W. D. Ross's claim that it is worse to harm than not to help someone. Also, the fact that the principle provided answers to nontraining cases that are consistent with Buchanan and Brock's judgments offers preliminary support for our hypothesis that decision principles discovered from some cases, using our method, enable a machine to determine the ethically acceptable action in other cases as well.

Conclusion

We have argued that machine ethics is an important new field of artificial intelligence and that its goal should be to create machines that are explicit ethical agents. We have done preliminary work to show—through our proof of concept applications in constrained domains—that it may be possible to incorporate an explicit ethical component into a machine. Ensuring that a machine with an ethical component can function autonomously in the world remains a challenge to researchers in artificial intelligence who must further investigate the representation and determination of ethical principles, the incorporation of these ethical principles into a system's decision procedure, ethical decision making with incomplete and uncertain knowledge, the explanation for decisions made using ethical principles, and the evaluation of systems that act based upon ethical principles.

Of the many challenges facing those who choose to work in the area of machine ethics, foremost is the need for a dialogue between ethicists and researchers in artificial intelligence. Each has much to gain from working together on this project. For ethicists, there is the opportunity of clarifying—perhaps even discovering—the fundamental principles of ethics. For AI researchers, convincing the general public that ethical machines can be created may permit continued support for work leading to the development of autonomous intelligent machines—machines that might serve to improve the lives of human beings.

Acknowledgements

This material is based upon work supported in part by the National Science Foundation grant number IIS-0500133.

Notes

1. See plato.stanford.edu/entries/logic-deontic.

2. A demonstration of MedEthEx is available online at www.machineethics.com.

3. For example, see www.ot.toronto.ca/iatsl/projects/medication.htm.

References

Allen, C.; Varner, G.; and Zinser, J. 2000. Prolegomena to Any Future Artificial Moral Agent. *Journal of Experimental and Theoretical Artificial Intelligence* 12(2000): 251–61.

Anderson, M., and Anderson, S., eds. 2006. Special Issue on Machine Ethics. *IEEE Intelligent Systems* 21(4) (July/August).

Anderson, M.; Anderson, S.; and Armen, C., eds. 2005a. Machine Ethics: Papers from the AAAI Fall Symposium. Technical Report FS-05-06, Association for the Advancement of Artificial Intelligence, Menlo Park, CA.

Anderson, M.; Anderson, S.; and Armen, C. 2005b. Toward Machine Ethics: Implementing Two Action-Based Ethical Theories. In Machine Ethics: Papers from the AAAI Fall Symposium. Technical Report FS-05-06, Association for the Advancement of Artificial Intelligence, Menlo Park, CA.

Anderson, M.; Anderson, S.; and Armen, C. 2006a. An Approach to Computing Ethics. *IEEE Intelligent Systems* 21(4): 56–63.

Anderson, M.; Anderson, S.; and Armen, C. 2006b. MedEthEx: A Prototype Medical Ethics Advisor. In *Proceedings of the Eighteenth Conference on Innovative Applications of Artificial Intelligence*. Menlo Park, CA: AAAI Press.

Anderson, S. L. 1995. Being Morally Responsible for an Action Versus Acting Responsibly or Irresponsibly. *Journal of Philosophical Research* 20: 453–62.

Asimov, I. 1976. The Bicentennial Man. In *Stellar Science Fiction 2,* ed. J.-L. del Rey. New York: Ballatine Books.

Baral, C. 2003. *Knowledge Representation, Reasoning, and Declarative Problem Solving*. Cambridge, UK: Cambridge University Press.

Beauchamp, T. L., and Childress, J. F. 1979. *Principles of Biomedical Ethics*. Oxford, UK: Oxford University Press.

Bentham, J. 1907. *An Introduction to the Principles and Morals of Legislation*. Oxford: Clarendon Press.

Bringsjord, S.; Arkoudas, K.; and Bello, P. 2006. Toward a General Logicist Methodology for Engineering Ethically Correct Robots. *IEEE Intelligent Systems* 21(4): 38–44.

Brody, B. 1988. *Life and Death Decision Making*. New York: Oxford University Press.

Buchanan, A. E., and Brock, D. W. 1989. *Deciding for Others: The Ethics of Surrogate Decision Making,* 48–57. Cambridge, UK: Cambridge University Press.

Capek, K. 1921. R.U.R. In *Philosophy and Science Fiction,* ed. M. Phillips. Amherst, NY: Prometheus Books.

Clarke, A. C. 1968. *2001: A Space Odyssey.* New York: Putnam.

Damasio, A.R. 1994. Descartes' Error: Emotion, Reason, and the Human Brain. New York: G. P. Putnam.

Dennett, D. 2006. Computers as Prostheses for the Imagination. Invited talk presented at the International Computers and Philosophy Conference, Laval, France, May 3.

Dietrich, E. 2006. After the Humans Are Gone. Keynote address presented at the 2006 North American Computing and Philosophy Conference, RPI, Troy, NY, August 12.

Ganascia, J. G. 2007. Using Non-Monotonic Logics to Model Machine Ethics. Paper presented at the Seventh International Computer Ethics Conference, San Diego, CA, July 12–14.

Gazzaniga, M. 2006. *The Ethical Brain: The Science of Our Moral Dilemmas*. New York: Harper Perennial.

Guarini, M. 2006. Particularism and the Classification and Reclassification of Moral Cases. *IEEE Intelligent Systems* 21(4): 22–28.

Horty, J. 2001. *Agency and Deontic Logic*. New York: Oxford University Press.

Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence
July, 2007 Vancouver, British Columbia, Canada
2 vols., references, index, illus., ISBN 978-1-57735-323-2
www.aaaipress.org

Joy, B. 2000. Why the Future Doesn't Need Us. *Wired Magazine* 8(04) (April).

Kant, I. 1785. *Groundwork of the Metaphysic of Morals,* trans. by H. J. Paton (1964). New York: Harper & Row.

Lavrec, N., and Dzeroski, S. 1997. *Inductive Logic Programming: Techniques and Applications*. Chichester, UK: Ellis Horwood.

Mappes, T. A., and DeGrazia, D. 2001. *Biomedical Ethics*, 5th ed., 39–42. New York: McGraw-Hill.

McLaren, B. M. 2003. Extensionally Defining Principles and Cases in Ethics: An AI Model. *Artificial Intelligence Journal* 150(1–2): 145–1813.

Moor, J. H. 2006. The Nature, Importance, and Difficulty of Machine Ethics. *IEEE Intelligent Systems* 21(4): 18–21.

Onishi, N. 2006. In a Wired South Korea, Robots Will Feel Right at Home. *New York Times,* April 2, 2006.

Picard, R. W. 1997. Affective Computing. Cambridge, MA: The MIT Press.

Pojman, L. J. 1996. The Case for Moral Objectivism. In *Do the Right Thing: A Philosophical Dialogue on the Moral and Social Issues of Our Time,* ed. F. J. Beckwith. New York: Jones and Bartlett.

Powers, T. 2006. Prospects for a Kantian Machine. *IEEE Intelligent Systems* 21(4): 46–51.

Rawls, J. 1951. Outline for a Decision Procedure for Ethics. *The Philosophical Review* 60(2): 177–197.

Ross, W. D. 1930. *The Right and the Good.* Oxford: Clarendon Press.

Rzepka, R., and Araki, K. 2005. What Could Statistics Do for Ethics? The Idea of a Common Sense Processing-Based Safety Valve. In Machine Ethics: Papers from the AAAI Fall Symposiu. Technical Report FS-05-06, Association for the Advancement of Artificial Intelligence, Menlo Park, CA.

Turing, A. M. 1950. Computing Machinery and Intelligence. *Mind* LIX(236): 433–460.

Michael Anderson is an associate professor of computer science at the University of Hartford, West Hartford, Connecticut. He earned his Ph.D. in computer science and engineering at the University of Connecticut. His interest in further enabling machine autonomy brought him first to diagrammatic reasoning where he cochaired Diagrams 2000, the first conference on the topic. This interest has currently led him, in conjuction with Susan Leigh Anderson, to establish machine ethics as a bona fide field of study. He has cochaired the AAAI Fall 2005 Symposium on Machine Ethics and coedited an *IEEE Intelligent Systems* special issue on machine ethics in 2006. His research in machine ethics was selected for IAAI as an emerging application in 2006. He maintains the machine ethics website (www.machineethics.org) and can be reached at anderson@hartford.edu.

Susan Leigh Anderson, a professor of philosophy at the University of Connecticut, received her Ph.D. in philosophy at UCLA. Her specialty is applied ethics, most recently focusing on biomedical ethics and machine ethics. She has received funding from NEH, NASA, and NSF. She is the author of three books in the Wadsworth Philosophers Series, as well as numerous articles. With Michael Anderson, she has presented work on machine ethics at national and international conferences, organized and cochaired the AAAI Fall 2005 Symposium on Machine Ethics, and coedited a special issue of *IEEE Intelligent Systems* on machine ethics (2006). She can be contacted at Susan.Anderson@uconn.edu.

View publication stats